Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 55 (2015) 503 – 513

Procedia Computer Science

Information Technology and Quantitative Management (ITQM 2015)

Design scheme for spatial database of climatic and environmental variables in Mexico, integrating Big Data Technology.

M. López[a], S. Couturier[b], K. Barrera[a,b,*]

[a] Laboratorio de Análisis Geoespacial, Instituto de Geografía, UNAM, Circuito Exterior, Cd. Universitaria, C.P.04510, México, D. F.
[b] *Laboratorio de Análisis Geoespacial Instituto de Geografía, UNAM, Circuito Exterior, Cd. Universitaria, C.P.04510, México, D. F.*
[a,b,] *Laboratorio de Análisis Geoespacial Instituto de Geografía, UNAM, Circuito Exterior, Cd. Universitaria, C.P.04510, México, D. F.*

Abstract

Big Data technology offers new and potentially more effective solutions to the continuing challenge of how large raster data sets may be effectively stored and retrieved. Big Data technology not only enables the integration of diverse information sources, but also supports the management and storage of large data sets. As such, it offers an alternative approach to existing spatial database technology. This paper proposes a design scheme that uses Big Data technology as a means to store for retrieval large volumes of raster data forming a spatial database of climatic and other environmental data for Mexico.

Peer-review under responsibility of the Organizing Committee of ITQM 2015

Keywords: Spatial Database; BigData; Raster; PMI; AVHRR; Scheme

1. Introduction

The problem of how to effectively handle large volumes of spatial data is not new. For decades [1] have implemented several methodologies for integrating the storage and retrieval of both vector and raster data. One of the first commercial relational database companies to address this problem was ORACLE [1], developed an extended data model with new data types that more effectively encapsulated the nature and properties of spatial information. Spatial indexes were also implemented which reduced response time for queries on both types of spatial data and improved the performance of these databases. In contrast, GIS companies such as ESRI created geo-databases, another approach thought capable of handling large volumes of spatial data [2]. Both approaches sought to deliver scalable solutions that allowed major clients such as national mapping organizations or utility companies to effectively deal with the challenges of maintaining large geographical data sets.

Various Open Source software has also been extended and new tools created to implement large spatial

Tel.: 56-22-42-40 ext. 45514; fax: 56-16-21-45
E-mail address:marco@igg.unam.mx.

Peer-review under responsibility of the Organizing Committee of ITQM 2015
doi:10.1016/j.procs.2015.07.026

databases, for example PostgreSQL, MySQL, Cassandra and MongoDB [3]. These various solutions have explored different means to store both non-spatial attribute information and vector and raster forms of spatial data in a single unified data model. Currently in the research field of large spatial databases the challenge is to both store the data sets with high physical compaction and at the same time support queries upon the different data types in an integrated and seamless manner. This implementation work is built upon and facilitated by new international standards concerning both spatial information and for managing database projects, such as ISO-TC211 [4] and the IEEE-1490-2011 [5].

2. Background

Since January 1996 the Institute of Geography of UNAM in Mexico City has operated a Receiving Station for Satellite Images (ERISA) that incorporates a receiving antenna and three servers for storing and processing these data.

The images received are:

Table 1. Received Images.

Satellite	Image	Periodicity	Size (Mbytes)
TIROS-NOAA [6]	NOAA-12 (Historical)	1x24 Diurnal	75 Mbytes
TIROS-NOAA	NOAA-12 (Historical)	1x24 Nocturnal	75 Mbytes
TIROS-NOAA	NOAA-14 (Historical)	1x24 Diurnal	75 Mbytes
TIROS-NOAA	NOAA-14 (Historical)	1x24 Nocturnal	75 Mbytes
TIROS-NOAA	NOAA-15 (Historical)	1x24 Diurnal	75 Mbytes
TIROS-NOAA	NOAA-15 (Historical)	1x24 Nocturnal	75 Mbytes
TIROS-NOAA	NOAA-16 (Historical)	1x24 Diurnal	75 Mbytes
TIROS-NOAA	NOAA-16 (Historical)	1x24 Nocturnal	75 Mbytes
TIROS-NOAA	NOAA-18	1x24 Diurnal	75 Mbytes
TIROS-NOAA	NOAA-18	1x24 Nocturnal	75 Mbytes
TIROS-NOAA	NOAA-19	1x24 Diurnal	75 Mbytes
TIROS-NOAA	NOAA-19	1x24 Nocturnal	75 Mbytes
GOES [7]	GOES	96x24	20 Mbytes

Description of requirements

The ERISA system receives daily a large amount of information from the receiving antenna of satellite images TERASCAN. This information is stored on magnetic tape for later use. However, the recovery process from tape for subsequent image processing is very slow and currently there is only an online capability to display the latest set of images received in "gif" format.

The present priority on archiving of received imagery and the limitation of working with mainly offline storage means that it is not possible to routinely carry out further processing of images, for example to derive secondary image information products, such as climatic and environmental variables.

A new system, capable of not only maintaining the past ERISA archive of terabytes of data, but which can also support rapid retrieval and more online and interactive processing and analysis of imagery is necessary.

2.1 Storage Needs

Table 2. AVHRR 1996 to 2010

Calculations from 1996 until 2010	AVHRR	Mbytes	
Size in Megabyte per recepcion day:		900	
Size in Megabyte per month		27000	
Size in Megabyte per year		324000	
Size in Gigabyte		316.4	
Size in Gigabyte per 14 years (1996-2010)		4429.68	4.32 Terabyte

Table 3. AVHRR 2010 to 2013

Calculations of 2011 to 2014	AVHRR	Mbytes	
Size in Megabyte per day		300	
Size in Megabyte per month		9000	
Size in Megabyte per year		108000	
Size in Gigabyte per year		105.46	
Size in Gigabyte per 3 years		316.4	0.3 Terabyte

The size of the storage system required to store the existing image archive is 5 Terabyte.

In order to continue to manage future received information for 15 years or more, an additional 5 Terabytes is needed.

The total storage required for this item is therefore estimated at 10 Terabyte

Table 4. GOES 1996 to 2010

Data received from 1996 until 2010	Mbytes	
Size in Megabyte per hour of reception	80	
Size in Megabyte per day of reception	1920	
Size in Megabyte per month	57600	
Size in Megabyte per year	691200	
Size in Gigabyte per year	675	
Size in Gigabyte per 15 years (96-2010)	9450	9.22 Terabytes

Table 5. GOES 2010 to 2011

Data received from 2010 to 2011	Mbytes	
Size in Megabyte per day	1920	
Size in Megabyte per month	57600	
Size in Megabyte per year	691200	
Size in Gigabyte per year	675	
Size in Gigabyte per 1 year	675	0.65 Terabyte

The storage required to store the 14 years of archived images is 10 TB. To manage information for a further 15 years, it is estimated that an additional 10 TB will be needed. This yields an estimate for the capacity of the new system of 20 TB.

Taken, together, the above two estimates reveal that the mass storage system that will serve as a repository for the database must have a minimum capacity of 30 Terabytes.

3. **Approach for implementing the system**

While the physical creation of such systems is not a simple task in itself, it is widely understood that the introduction of any new spatial database or GIS needs careful planning. Workers such as Tomlinson [2], Aronoff [8] and Huxhold [9] have devised planning methods specifically for implementing GIS projects that had been built up from experience to reduce the risks of project failure. These methodologies are designed to assesss the information needs and the business environment for the client organization and see such implementations as social as well as technical processes. More recently, it has become common practice to use more industry standard project management methodologies to manage the implementation of GIS/spatial database projects [10], as these tools are deployed in many organizations as part of enterprise solutions.

It is clear, that information is a key part of the system and of the success depends on the initial analysis of it. Another point is the development scheme to be used in the construction of the database. In this project for the construction of a storage system of satellite images AVHRR (Advanced Very High Resolution Radiometer) the five process groups posed by the Project Management Institute (PMI) and the ten areas of knowledge that integrates need to be taken into account [11].

Investigations have been conducted in recent years [12] have demonstrated the increased use of PMI project management compared to other standards such as PRINCE2 (Projects In Controlled Environments) and PMMM (Project Management Maturity Model). While the establishment of its use is not easy, a crucial point is the way in which knowledge is transmitted developers, is why in this scheme will be presented in a simple way to make it easy implementation.

3.1 Initial Planning

Following the approach advocated by the PMI [13]:

I. Get user requirements.
II. Designate the system requirements
III. Perform feasibility study

IV. Set the system architecture
V. Establish the system design

3.2 Integration with project management

Once you have the information needed for the project should be taken as a guide the following groups of processes established by PMI:

- Initiation
- Planning
- Execution
- Monitoring and control
- Close

The sequence of these groups is the following processes:

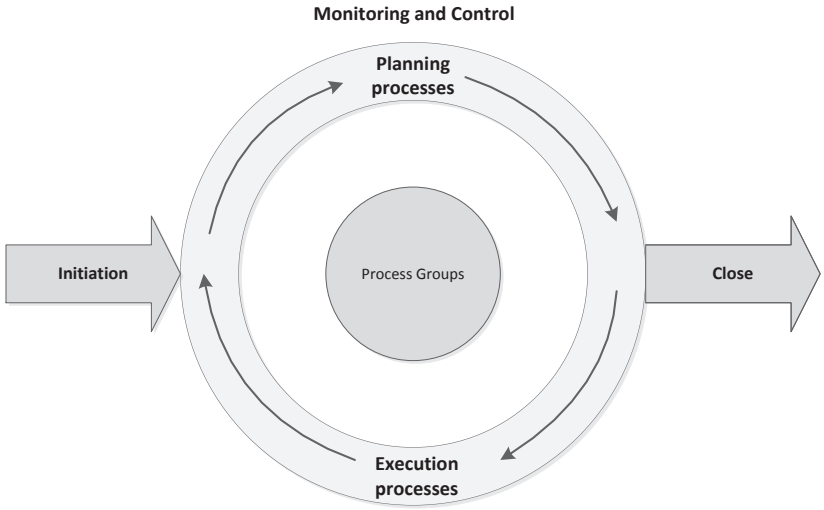

Fig 1. Processes established by PMI [9]

The sequence of processes above is similar to the model of continuous improvement known as PDCA (Plan, Do, Check, Act) as set out by Edward Deming [14].

The proposed implementation scheme is split into two main stages:

I. Preliminary Steps
II. System Development

I. Preliminary Steps

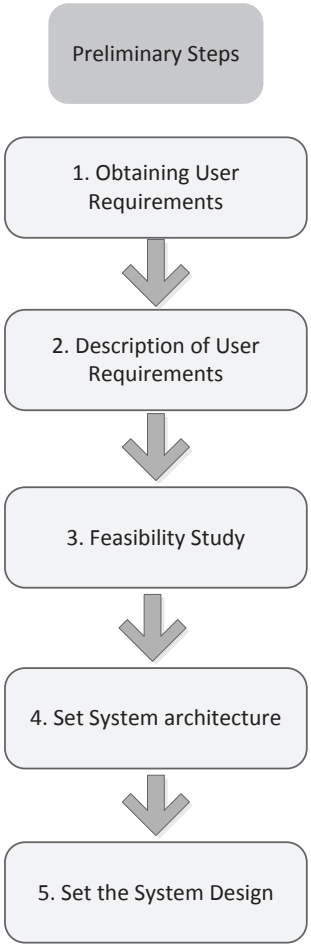

Fig. 2. Preliminary Steps

The omission of this step leads to the low probability of project success, as they have before developing a detailed analysis of the system, leading to errors and misunderstandings between the members of the working group and therefore is an important part in the proposed scheme.

II. System Development

At this stage the groups of processes by which they are installed and configured both hardware and software and integration of data in the system will be made.

The processes of planning and design to be used for the new system are illustrated below:

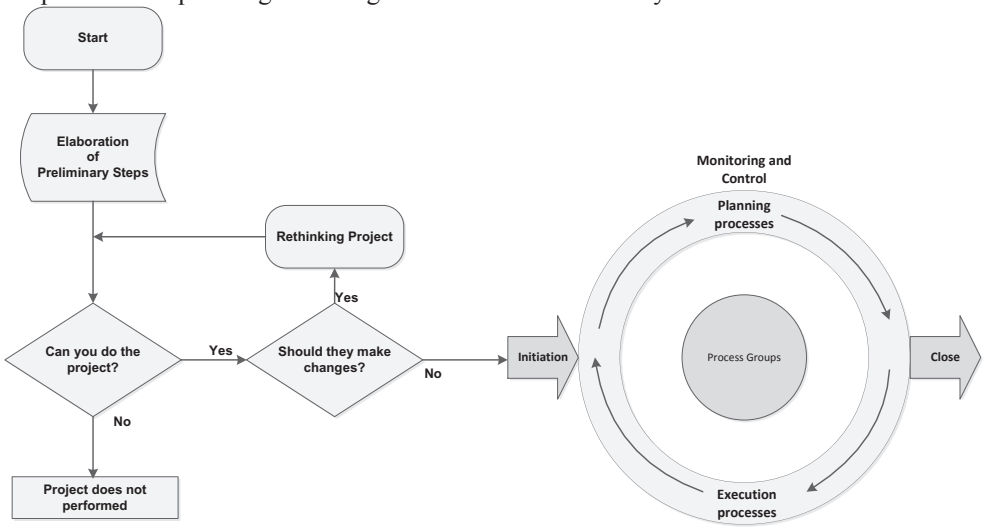

Fig 3. Approach to be used to system implementation

4. **Implementation**

I. Preliminary Steps

1. Obtaining user requirements

The requirements:

a) Storage of archive information and further 15 years of data
b) Fast Recovery of stored data
c) Agile Consultations
d) Consultation by Internet browsers
e) Ability to associate other images and docs e.g. pictures
f) Retrieval of metadata: (e.g. sensor, date and time) for each image

2. Designation of user requirements
3. Feasibility study, consisting of:
   a) Operational testing
   b) Technical
   c) Economic (cost benefit)

4. System Architecture

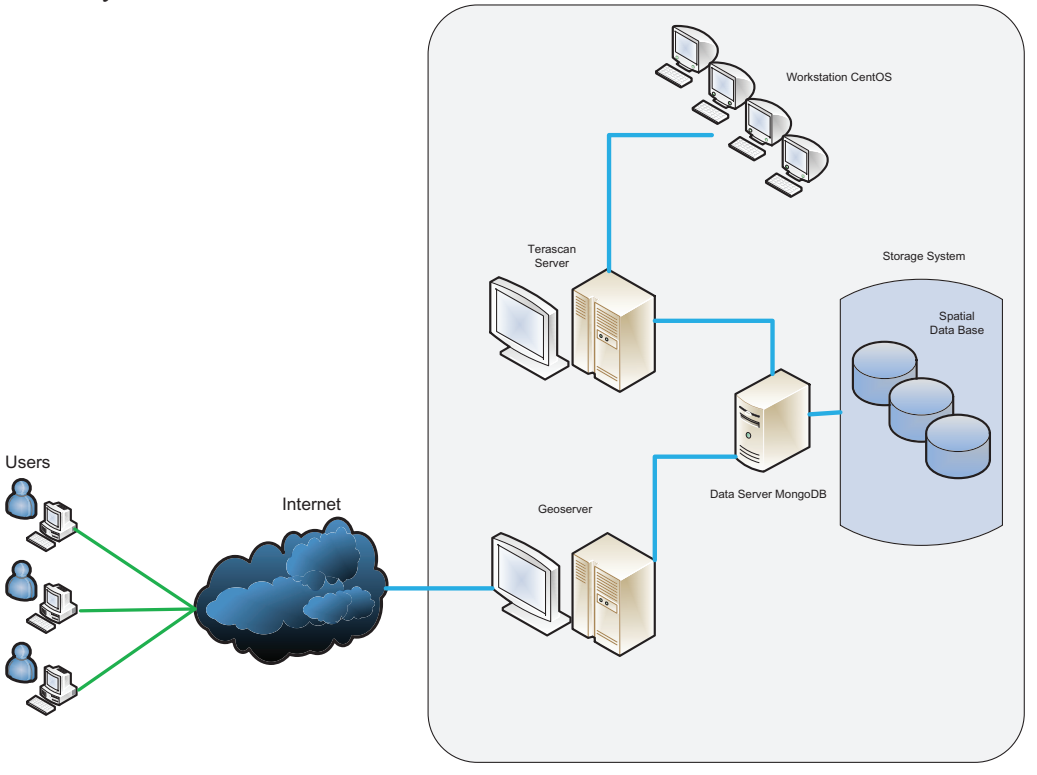

Fig 4. System Architecture

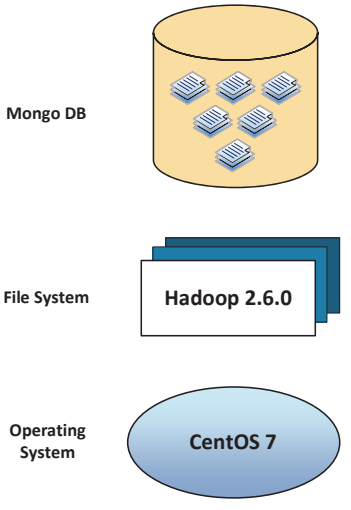

Fig 5. System Diagram

5. System Design

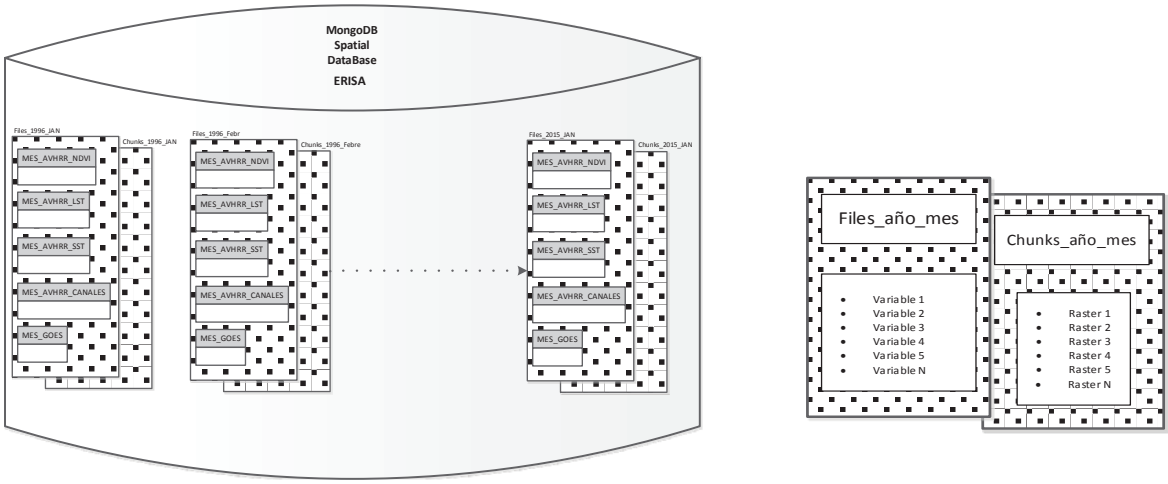

Fig. 6. File Organization in MongoDB

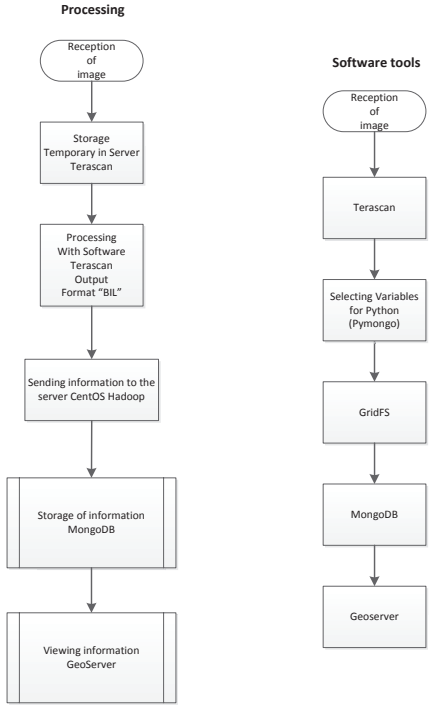

Fig 7. Flow of data and processing

Table 6. Air Pollution and Climatic Variables

Satellite	Sensor	Product	Variable
NOAA	AVHRR 1, AVHRR 2 o AVHRR 3	Registered	Visible Albedo
			IRC Albedo
			IRM Temperature
			IRT1 Temperature
			IRT2 Temperature
NOAA	AVHRR	Sea Surface Temperature	Temperature
NOAA	AVHRR	Normalized Difference Vegetation Index	
NOAA	AVHRR	Soil Temperature	
NOAA	AVHRR	Cloud Top	
NOAA	AVHRR	Cloud Mask	
NOAA	AVHRR	Visibility	
NOAA	AVHRR	Water Vapour	
NOAA	TOVS	Surface Pressure	
NOAA	TOVS	Precipitation Water	
NOAA	TOVS	Wind speed	
NOAA	TOVS	Albedo	Surface Radiation Budget
NOAA	TOVS	Skin Temperature	Air Temperature
NOAA	TOVS	Cleared Cloudy	Cloud properties
NOAA	TOVS	Atmospheric Temperature	
NOAA	TOVS	Wind Speed	
NOAA	ATOVS	Dew point	
NOAA	ATOVS	Ozone	
NOAA	ATOVS	Water Vapour	
NOAA	ATOVS	Solar Irradiance	

II. System Development

Considering the IEEE 1490-2011 standard for the adoption of PMI in developing software projects and systems engineering, the following activities were established:

A) Configuration CentOS operating system to install the MongoDB database manager data
B) Installing and configuring Hadoop [15]
C) Install and configure MongoDB
D) Construction of the elements according to the structure of Spatial Database
E) Creating and configure users
F) Integration of information reference Geoserver
G) Installation and configuration of GeoServer

5. Conclusion

This paper has described the design and discussed the proposed stages in the implementation of a system for storing, managing and distributing both archive and future images to be generated by the ERISA system, located at the Institute of Geography at UNAM. The proposed system provides a solution to some common problems of storing the typically large volumes of spatial data produced by receiving satellite data.

The new requirements for more interactive browsing of imagery and processing of image data to derive additional climatic and environmental variables implies an increase in both the storage capacity and functionality needed from the spatial database. The database Mongo DB has been identified for system prototyping, because it can handle large data volumes and also provides compatibility with other software tools such as Python and GeoServer, through which algorithms can be implemented for processing the imagery and the display of the results.

The proposed architecture hosts the new database that will be responsible for receiving and processing the satellite images on a separate server, as new processes will generate a considerable additional workload that would be expected to degrade the performance of the existing "Terascan" server for basic archiving functions.

The implementation of the new database using MongoDB is not expected to be a straightforward task, and various stages of testing and refinement of the database parameters are expected to be required in order to achieve satisfactory performance.

6. References

[1] Yeung, A. & Brent, G. Spatial Database Systems. Canada: Springer. 2007.

[2] Tomlinson, R. Pensando en SIG. 3rd ed. EUA: ESRI Press. 2007.

[3] Plugge, E., Membrey, P., & Hawkins, T. The definitive guide to Mongo DB. The NoSQL Database for cloud and desktop computing. USA: Apress. 2010.

[4] ISO/TC 211. Report from stage 0Project 1950 Geographic Information Ontology. 2009.

[5] IEEE. Guide-Adoption of the Project Management Institute (PMI) Standard. A guide to the Project Management Body of Knowledgement (PMBOK Guide) 4ta ed. USE: IEEE Computer Society. 2011.

[6] http://www.lib.noaa.gov/collections/TIROS/tiros.html

[7] http://www.goes.noaa.gov/

[8] Aronoff, S. Geographic Information Systems: A Management Perspective. WDL Publications. Ottawa, Canada. 1990.

[9] Huxhold, W. E. An Introduction to Urban Geographic Information Systems. New York: Oxford University Press. 1991.

[10] Pant, I., & Bassam, B. Project management education: The human skills imperative. En: International Journal of Project Management 2008;26.

[11] PMI. PMBOK. Fundamentos para la dirección de proyectos. 5ed EUA: PMI. 2013.

[12] Ahlemann, F., Teuteberg, F., & Vogelsang, K. Project management standards – Diffusion and application in Germany and Switzerland. International Journal of Project Management 2009; 27:292–303

[13] PMI. PMBOK. Fundamentos para la dirección de proyectos. 4ed EUA: PMI. 2008.

[14] Naaranoja, M., Päivi, H., & Heikki, L. Strategic management tools in projects case construction Project. International Journal of Project Management 2007; 25: 659–665

[15] Hows, D., Membrey, P. & Plugge, E. Mongo DB basics. USA: Apress. 2014.